Statistische und stochastische Grundlagen

Vorlesungsmitschrieb zum Modul an der Universität Stuttgart

Simon König

Inhaltsverzeichnis

1

Kapitel 1: Statistik

1: Grundbegriffe

1.1 Grundbegriffe der Statistik

Statistische Einheit Objekte die erfasst werden und an denen die interessierenden Größen erfasst werden

Grundgesamtheit Menge aller für die Fragestellung relevanten statistischen Einheiten

Teilgesamtheit Teilmenge der Grundgesamtheit

Stichprobe Tatsächlich untersuchte Teilmenge der Grundgesamtheit

Merkmal, Variable Größe von Interesse

Merkmalsausprägung, Wert Konkreter Wert des Merkmals für eine bestimmte statistische Einheit

1.2 Charakterisierung der Merkmale

diskret Merkmale, die nur endlich viele oder abzählbar unendlich viele Ausprägungen annehmen sind diskret.

stetig Merkmale, die Werte aus einem Intervall annehmen können heißen stetig.

quasi-stetig Merkmale, die sich nur diskret messen lassen aber aufgrund einer sehr feinen Abstufung wie stetige Merkamle behandelt werden können.

Die Ausprägungen eines stetigen Merkmals lassen sich immer so zusammenfassen, dass es als diskret angesehen werden kann. Die Ausprägungen heißen dann gruppiert oder klassiert.

1.3 Skalen

Zusätzlich zur Charakterisierung der Merkmale werden diese anhand ihres Skalenniveaus unterschieden.

Nominalskala Wenn die Ausprägungen Namen oder Kategorien sind, die den Einheiten zugeordnet werden heißt das Merkmal *nominalskaliert*. Beispielsweise Geschlecht oder Verwendungszweck.

Ordinalskala Merkmale mit Ausprägungen zwar mit Ordnung, bei denen allerdings ein Abstand der Merkamale nicht interpretier- oder vergleichbar ist heißen *ordinalskaiert*. Ein Beispiel hierfür wären Schulnoten.

Kardinalskala Ein kardinalskaliertes Merkmal wird oft auch metrisch bezeichnet. Hierbei sind die Abstände der Ausprägungen interpretierbar und zusätzlich ist ein sinnvoller Nullpunkt der Skala festgelegt oder bestimmbar.

Auf Basis dieser Skalenmerkmale nennt man Merkmale mit endlich vielen Ausprägungen, die höchstens ordinalskaliert sind *qualitative* oder *kategoriale Merkmale*. Diese geben eine Qualität aber nicht ein Ausmaß wieder.
Geben die Ausprägungen jedoch eine Intensität oder Ausmaß wieder so spricht man von *quantitativen Merkmalen*. Alle Messungen mit Zahlenwerten stellen Ausprägungen quantiativer Merkmale dar. Ein kardinalskaliertes Merkmal ist stets quantitativ.

1.4 Datengewinnung, Datenerhebung

S.18 ff

2: Verteilungen und ihre Darstellungen

2.1 Häufigkeiten

Als *Urliste* bezeichnet man die Menge der Merkmale X der Untersuchungseinheiten $U = \{x_1, \ldots, x_n\}$. Die *auftretenden Ausprägungen* von X sind die Werte $\{a_1, \ldots, x_n\} \subseteq \{x_1, \ldots, x_n\}, k \leq n$. Oftmals treten in einem großen Datensatz der Größe n nicht auch n verschiedene Werte x_i auf.
Damit definieren sich

Definition 2.1: Absolute Häufigkeit

Die absolute Häufigkeit einer auftretenden Ausprägung a in einer Urliste U ist

$$h(a) = |\{i \in \mathbb{N} \mid x_i = a, x_i \in U\}|.$$

Es gilt immer, dass die Summe aller absoluten Häufigkeiten gleich der Datensatzgröße ist

$$\sum_{i=1}^{n} h(a_i) = |U|.$$

Die absolute Häufigkeitsverteilung ist dargestellt durch die Folge von Werten

$$h_1, \ldots, h_k = h(a_i), \ldots, h(a_k)$$

Definition 2.2: Relative Häufigkeit

Die relative Häufigkeit einer auftretenden Ausprägung a in einer Urliste U ist

$$f(a) = \frac{h(a)}{|U|}.$$

Es gilt ähnlich wie bei der absoluten Häufigkeit für die Summe

$$\sum_{i=1}^{n} f(a_i) = 1.$$

Eine grafische Darstellung einer Häufigkeitsverteilung nennt man ein *Histogramm*. Bei Histogrammen ist auf die Flächentreue zu achten, das bedeutet, dass der Flächeninhalt der aufgetragenen Rechtecke proportional (oder gleich) zu h_j oder f_j ist. So kann das menschliche Auge die Verteilung besser wahrnehmen.
Hat das Histogramm einer Verteilung nur einen deutlich erkennbaren Hochpunkt (Gipfel), heißt sie *unimodal*. Treten mehrere Gipfel auf nennt man die Verteilung *multimodal*. Bei zwei Gipfeln spricht man von einer *bimodalen* Verteilung.

Man nennt eine Verteilung *symmetrisch*, wenn es eine Symmetrieachse gibt, sodass die rechte und linke Hälfte der Verteilung annähernd zueinander spiegelbildlich sind. Eine Verteilung heißt *schief*, wenn sie deutlich unsymmetrisch ist. Sie heißt dann *linkssteil oder rechtsschief*, wenn der überwiegende Anteil von Daten linksseitig konzentriert ist. Dann steigt die Verteilung links deutlich steiler ab als rechts. Entsprechend *rechtssteile oder linksschiefe* Verteilungen.

2.2 Kumulierte Häufigkeiten

Die kumulierten Häufigkeitsverteilungen geben an, wie viele Datenpunkte der Urliste, beziehungsweise welcher Anteil der Daten unterhalb einer Schranke liegen. Um diese Aussage sinnvoll zu beantworten ist zumindest eine Ordinalskala nötig.

Definition 2.3: Absolute kumulierte Häufigkeitsverteilung

Die absolut kumulierte Häufigkeitsverteilung ist die Funktion

$$H(x) = \sum_{i:a_i \leq x} h_i.$$

Definition 2.4: Relative kumulierte Häufigkeitsverteilung

Die relative kumulierte Häufigkeitsverteilung oder auch empirische Verteilungsfunktion ist

$$F(x) = \sum_{i:a_i \leq x} f_i.$$

Die kumulierten Häufigkeitsverteilungen sind monoton wachsende, Treppenfunktionen, die an den Sprungstellen rechtsseitig stetig sind.

2.3 Gruppierung

Sind alle auftretenden Ausprägungen Elemente eines Interfalls $[a, b]$, lässt sich dieses in gleich große Klassen der Größe d unterteilen.
Eine Klassifizierung ist allgemein

$$[a, c_1), \ldots, [c_i, c_{i+1}), [c_{i+1}, c_{i+2}), \ldots \quad \forall i : c_{i+1} - c_i = d$$

Klassifizierte Daten sind i.A. einfacher zu interpretieren als große Mengen von Daten, die sich nur wenig voneinander unterscheiden.
Der maximale Fehler bei der Klassifizierung ist die halbe Klassengröße.

2.4 Lagemaße

Lagemaße helfen beim Vergleich verschiedener Eigenschaften, bzw dem Vergleich verschiedener statistischer Einheiten mit einer gemeinsamen Eigenschaft.

Definition 2.5: Lagemaß

Ein *Lagemaß* ist eine Abbildung $L : \mathbb{R}^n \to \mathbb{R}$ mit der Eigenschaft

$$L(x_1 + a, \ldots, x_n + a) = L(x_1, \ldots, x_n) + a \quad \forall a, x_i \in \mathbb{R} \ (1 \leq i \leq n)$$

Ein Lagemaß beschreibt das Zentrum einer Verteilung.

Beispiele für Lagemaße sind

2.4.1 Arithmetisches Mittel

Das arithmetische Mittel ist nur für quantitative Merkmale sinnvoll. Es berechnet sich durch

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} (a_i \cdot f_i)$$

aus Rohdaten beziehungsweise aus den Häufigkeitsdaten.
Mit dem arithmetischen Mittel gilt die sogenannte *Schwerpunkteigenschaft*

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0.$$

Unter einer linearen Transformation $x \mapsto ax + b$ verhält sich das arithmetische Mittel analog

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} (ax_i + b) = \frac{a}{n} \sum_{i=1}^{n} x_i + b = a\overline{x} + b$$

Wie aus der Formel erkennbar, ist das arithmetische Mittel extrem empfindlich gegen Ausreißer. Dafür wurden die folgenden Mittel eingeführt.

Das getrimmte Mittel Um Ausreißer weniger stark ins Gewicht fallen zu lassen wird der Datensatz absichtlich verkleinert. Beim getrimmten Mittel aus einer sortiert vorliegenden Liste von Daten werden zum Beispiel die oberen und unteren 5% der Daten abgeschnitten, damit fallen auch eventuelle Ausreißer raus. Die Datensatzgröße bleibt jedoch nicht erhalten.

Das winsorisierte Mittel Ähnlich wie beim getrimmten Mittel wird der Datensatz beim winsorisierten Mittel von oben und unten herein bearbeitet. Anstatt Daten zu löschen werden beispielsweise die oberen 5% durch den nächstkleineren Wert ersetzt. Hierbei bleibt also die Datensatzgröße gleich.

2.4.2 Median

Der Median stellt ein robusteres Lagemaß als das arithmetische Mittel dar, er ist resistenter gegen Ausreißer im Datensatz. Für $x_1 \leq x_2 \leq \ldots \leq x_n$, also einen sortiert vorliegenden Datensatz ist der Median

$$x_{\text{med}} = \begin{cases} x_{\frac{n+1}{2}} & n \text{ ungerade} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & n \text{ gerade} \end{cases}$$

Benötigt eine Zahlenordnung, also eine Ordinalskala.
Der Modus verhält sich unter linearer Transformation $y = ax + b$ genauso wie das arithmetische Mittel $y_{\text{med}} = ax_{\text{med}} + b$.
Mindestens 50% der Daten sind kleiner oder gleich x_{med}, genauso sind mindestens 50% der Daten größer oder gleich dem Median x_{med}.

2.4.3 Modus

Der Modus ist die Ausprägung größter Häufigkeit $x_{\text{mod}} = a_i$ mit $h(a_i) = \max\{h(a) \mid a \in A\}$ wobei A die Menge aller vorkommenden Ausprägungen der Urliste ist. Der Modus ist dann eindeutig, wenn die Häufigkeitsverteilung ein eindeutiges Maximum besitzt.
Der Modus empfiehlt sich schon für nominalskalierte Daten.
Der Modus verhält sich unter linearer Transformation $y = ax + b$ genauso wie das arithmetische Mittel $y_{\text{mod}} = ax_{\text{mod}} + b$.

2.4.4 Geometrisches Mittel

Für eine Urliste $U = \{u_1, \ldots, u_n\}$ ist das geometrische Mittel definiert als

$$x_{\text{geom}} = \sqrt[n]{\prod_{i=1}^{n} u_i}.$$

Das geometrische Mittel wird z.B. bei der Berechnung des effektiven Jahreszinses verwendet, es stellt jedoch kein Lagemaß im engeren Sinne dar.

2.4.5 Harmonisches Mittel

Für eine Urliste $U = \{u_1, \ldots, u_n\}$ ist das harmonische Mittel definiert als

$$x_{\text{harm}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Genauso wie das geometrische Mittel zählt das harmonische nicht zu den Lagemaßen im engeren Sinne.

2.5 Lageregeln

Für symmetrische Verteilungen gilt $\overline{x} \approx x_{\text{med}} \approx x_{\text{mod}}$.
Für linkssteile Verteilungen $\overline{x} > x_{\text{med}} > x_{\text{mod}}$.
Und ebenso für rechtssteile Verteilungen $\overline{x} < x_{\text{med}} < x_{\text{mod}}$.

2.6 Streuungsmaße

Um eine Verteilung sinnvoll beschreiben zu können sind zusätzlich zu den Lagemaßen noch Aussagen über die Streuung der Daten um das Mittel nötig.

Definition 2.6: Streuungsmaß

Ein *Streuungsmaß* ist eine Abbildung $S : \mathbb{R}^n \to \mathbb{R}$ für die gilt

$$S(x_1 + a, \ldots, x_n + a) = S(x_1, \ldots, x_n) \quad \forall a, x_i \in \mathbb{R}(1 \leq i \leq n)$$

Ein Streuungsmaß stellt dar, wie weit gestreut Werte einer Verteilung um ein Mittel liegen.

2.6.1 Spannweite, Stichprobenspannweite

Die Stichproblenspannweite stellt dar, in welchem Bereich die Ausprägungen liegen, denn diese ist einfach

$$x_{\max} - x_{\min}.$$

2.6.2 Mittlere absolute Abweichung vom Median

Die mittlere absolute Abweichung vom Median stellt eine robustere Alternative zur Stichprobenvarianz dar. Sie ist für eine Urliste $U = \{x_1, \ldots, x_n\}$ definiert als

$$\frac{1}{n} \sum_{i=1}^{n} |x_i - x_{\text{med}}|$$

2.6.3 Quantile

Ein Quantil ist eine Kennzahl, die Daten nach einer relativen Häufigkeit trennt. So trennt das p-Quantil einer Verteilung die Daten som dass etwa $p \cdot 100\%$ der Daten darunter und $(1-p) \cdot 100\%$ darüber liegen. Damit ist der Median gerade das 50%-Quantil.

Definition 2.7: Quantile

Sei $U = \{x_1, \ldots, x_n\}$ eine geordnete Urliste, d.h. $x_1 \leq \ldots \leq x_j \leq \ldots \leq x_n$. Das p-Quantil x_p ist eine Ausprägung $x_p \in U$ für die gilt

$$\frac{\left|\left\{i \in \mathbb{N} \,\middle|\, x_i \leq x_p, x_i \in U\right\}\right|}{n} \geq p \text{ und } \frac{\left|\left\{i \in \mathbb{N} \,\middle|\, x_i \geq x_p, x_i \in U\right\}\right|}{n} \geq 1 - p$$

Das heißt es liegen $p\%$ der Daten unterhalb und $(1-p)\%$ der Daten oberhalb des p-Quantils. Sinnvoll berechnen lässt sich das p-Quantil durch die Formel

$$x_p = \begin{cases} \frac{1}{2}(x_{(n \cdot p)} + x_{(n \cdot p+1)}) & \text{, falls } n \cdot p \text{ ganzzahlig} \\ x_{(\lfloor n \cdot p \rfloor)+1} & \text{sonst} \end{cases}$$

Dabei nennt man das 25%-Quantil auch das *untere Quartil* und entsprechend das 75%-Quantil das *obere Quartil*.

Definition 2.8: Interquartilsabstand

Für metrische Merkamel ist der sogenannte *Interquartilsabstand (interquartile range)* die Distanz

$$d_Q = \text{IQR} = x_{0.75} - x_{0.25}.$$

Der IQR wird zum Beispiel beim Box-Plot verwendet.

Mit dem IQR können Zäune festgelegt werden, außerhalb derer sich höchstwahrscheinlich Ausreißer des Merkmals befinden. Ein Beispiel hierfür ist zum Beispiel der untere Zaun $z_u = x_{0.25} - 1.5 \cdot d_Q$ und entsprechend die Obergrenze $z_o = x_{0.75} + 1.5 \cdot d_q$, diese Werte werden wiederum beim Box-Plot verwendet.

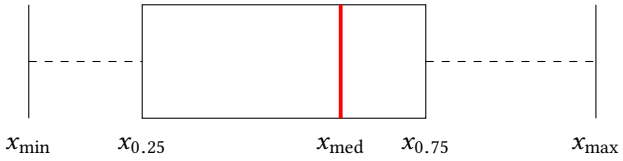
2.6.4 Fünf-Punkte-Zusammenfassung

Die Quartile, das Minimum, Maximum sowie der Median teilen den Datensatz in vier Teile, wobei jeder etwa ein Viertel der Merkmale enthält. Die Angabe dieser fünf Werte wird auch als Fünf-Punkte-Zusammenfassung bezeichnet.

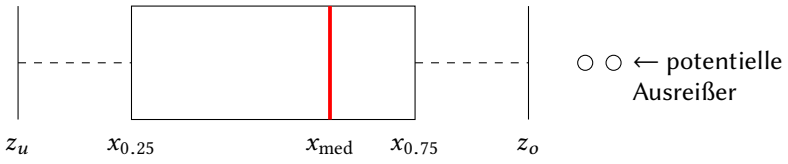
Box-Plot

Das Box-Plot ist eine einfache Art und Weise die Ausprägungen einer Verteilung zu visualisieren. Für den Box-Plot wird eine fünf-Punkte-Zusammenfassung, $(z_u, x_{0.25}, x_{\text{med}}, x_{0.75}, z_o)$ verwendet, wobei z_u und z_o beim modifizierten Box-Plot von der klassischen fünf-Punkte-Zusammenfassung abweichen können. Daran werden zwei Definitionen unterschieden, der „normale“ und der modifizierte Box-Plot.
Beim Box-Plot wird ein Rechteck zwischen den Quartilen gezeichnet, das x_{med} eingezeichnet als Linie oder Punkt beinhaltet. So sieht man dass sich 50% der Datenpunkte innerhalb der Box befinden. Die nach außen gezeichneten Linien geben an, wie weit die restlichen 50% der Datenpunkte gestreut liegen. Diese sogenannten Whiskers enden in Abhängigkeit von z_u bzw. z_o, diese Werte unterscheiden sich bei den beiden Definitionen.

Normaler Box-Plot Das Fünftupel besteht aus den Werten $(x_{\text{min}}, x_{0.25}, x_{\text{med}}, x_{0.75}, x_{\text{max}})$. Wobei x_{min} und x_{max} die kleinste und größte Ausprägung der Verteilung darstellen. So sind alle Werte in der Spannweite der Whiskers enthalten. Ein Box-Plot sieht dann wie folgt aus



Modifizierter Box-Plot (nach Fahrmeir) Der wichtigste Unterschied zum normalen Box-Plot ist dass, anstatt der minimalen und maximalen Werte für z_u, z_o ein Zaun gewählt wurde. So ist $z_u = x_{0.25} - 1.5 \cdot d_Q$ und $z_o = x_{0.75} + 1.5 \cdot d_Q$ wobei d_Q der Interquartilsabstand ist. Allerdings ist zu beachten dass die Whiskers von der größten/ kleinsten Ausprägung innerhalb des Zauns zur Box ausgehen. Liegen also beispielweise innerhalb des Bereichs $[z_u, x_{0.25}]$ keine Datenwerte, so existiert kein unterer Whisker. Datenpunkte, die außerhalb des Zauns liegen, werden mit Punkten dargestellt. Dies ist ein gutes Anzeichen für eventuelle Ausreißer.



2.6.5 Standardabweichung

Die Standardabweichung ist für eine Urliste $U = \{x_1, \ldots, x_n\}$ von metrischen Daten definiert als

$$\tilde{s} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\sum_{i=1}^{n}(a_i - \overline{x})^2 \cdot f_i}$$

wobei die a_i die Ausprägungen der Urliste sind und f_i die relative Häufigkeit der Ausprägung a_i ist.
Unter einer linearen Transformation $x \mapsto ax + b$ der Ausprägungen wird $\tilde{s}$ um $|a|$ gedehnt. Eine Verschiebung der Werte um b hat keine Auswirkung.

2.6.6 Variationskoeffizient

Der Variationskoeffizient ist eine maßstabsunabhängige Maßzahl für die Streuung, sie basiert auf der Standardabweichung und ist definiert als

$$v = \frac{\tilde{s}}{\overline{x}}, \quad \overline{x} > 0$$

2.6.7 Varianz, empirische Varianz

Die empirische Varianz ist das Quadrat der Standardabweichung $\tilde{s}^2$.

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Sowohl Standardabweichung als auch die Varianz sind nicht resistent, reagieren also sehr empfindlich auf Ausreißer.
Für eine Berechnung von Hand gilt der sogenannte Verschiebungssatz

Satz 2.9: Verschiebungssatz

Für jedes $c \in \mathbb{R}$ gilt

$$\sum_{i=1}^{n}(x_i - c)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - c)^2.$$

Damit gilt insbesondere mit $c = 0$ für die Varianz

$$\tilde{s}^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}.$$

Beweis:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - c)^2 &= \sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - c)^2 \\
&= \sum_{i=1}^{n}\left[(x_i - \overline{x})^2 + 2(x_i - \overline{x})(\overline{x} - c) + (\overline{x} - c)^2\right] \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + \underbrace{2(\overline{x} - c)\sum_{i=1}^{n}(x_i - \overline{x})}_{=0} + \sum_{i=1}^{n}(\overline{x} - c)^2 \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - c)^2
\end{aligned}$$

Lineare Transformation Mit einer linearen Abbildung der Ausprägungen $y_i = ax_i + b$ verhält sich die Varianz der Daten y_i

$$\tilde{s}_y^2 = a^2\tilde{s}_x^2 \text{ bzw. } \tilde{s}_y = |a|\tilde{s}_x$$

dies ergibt sich direkt aus den Formeln für die Varianz

$$\begin{aligned}\tilde{s}_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(ax_i + b - a\overline{x} - b)^2 \\ &= a^2\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = a^2\tilde{s}_x^2\end{aligned}$$

2.6.8 Stichprobenvarianz

Die Stichprobenvarianz stellt ein nicht resistentes Streuungsmaß dar. Sie ist für eine Urliste $U = \{x_1, \ldots, x_n\}$ definiert als

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}x_i^2 - n \cdot \overline{x}^2$$

2.7 Konzentrationsmaße

Konzentration geben Aufschluss über die Stärke der Konzentration von Daten.

2.7.1 Lorenzkurve

Ausgehend von geordneten Ausprägungen $x_1 \leq \ldots \leq x_n$ stellt die Lorenzkurve den Anteil der kumulierten relativen Merkmalssumme bezüglich dem Anteil der Merkmalsträger von der Grundgesamtheit dar.

Definition 2.10: Lorenzkurve

Die Lorenzkurve ergibt sich als Streckenzug durch die Punkte $(0,0), (u_1, v_1), \ldots, (1,1)$ wobei

$$u_j = \frac{j}{n} \quad \text{und} \quad v_j = \frac{\sum_{i=1}^{j} x_i}{\sum_{i=1}^{n} x_i}$$

ist.

Der Grundgedanke ist, darzustellen auf welchen Teil der Merkmalsträger welcher Anteil der Merkmalssumme zuruckgeht.
Die Lorenzkurve wächst immer monoton und konvex, d.h. sie wölbt sich nach unten.

2.7.2 Gini-Koeffizient

In der Lorenzkurve drückt sich Konzentration der Daten durch Entfernung von der ersten Winkelhalbierenden aus. Genau dies nutzt der Gini-Koeffizient G aus.
G ist gleich dem Verhältnis zwischen dem von Lorenzkurve und Diagonale eingeschlossenen Flächeninhalt und der Fläche unter der Winkelhalbierenden.

$$G = \frac{2\sum_{i=1}^{n} i x_i}{n\sum_{i=1}^{n} x_i} - \frac{n+1}{n} \quad G \in [0, \infty)$$

Dabei ist der minimale Wert des Gini-Koeffizienten $G_{\text{min}} = 0$ und das Maximum ist $G_{\text{max}} = \frac{n-1}{n}$.
Da der Wert des Gini-Koeffizienten mit von der Eingabegröße abhängnen kann, bietet sich der normierte Gini-Koeffizient für Vergleiche an

$$G^* = \frac{G}{G_{\text{max}}} = \frac{n}{n-1}G \quad G^* \in [0,1]$$

2.8 Schiefe und Wölbung

3: Multivariate

Bisher wurden nur eindimensionale Daten erfasst, nicht aber verschschiedene Merkmale in Zusammenhang gebracht und gemeinsam betrachtet oder miteinander verglichen.

3.1 Kontingenztafel

Die Kontingenztabelle eignet sich zu Darstellung der gemeinsamen Verteilung von zwei Diskreten Merkmalen mit relativ wenigen Ausprägungen.
Auf Basis der Ausprägungen $a_1, \dots, a_k$ des Merkmals X und $b_1, \dots, b_m$ für Y liegen in der Urliste die gemeinsamen Messwerte vor. Das heißt die Urliste besteht aus den Tupeln (a_i, b_j). Analog zum Eindimensionalen sind die absoluten Häufigkeiten h_{ij} definiert. Darauf aufbauend ebenfalls völlig analog die relativen Häufigkeiten f_{ij}.

Kontingenztafel der absoluten Häufigkeiten Die aus diesen Werten entstehende Tafel heißt $(k \times m)$-Kontingenztafel der absoluten Häufigkeiten. Sie enthält neben den Häufigkeitsdaten zusätzlich noch die Spalten- beziehungsweise Zeilensummen der Werte.

	b_1	$\cdots$	b_m	
a_1	h_{11}	$\cdots$	h_{1m}	$h_{1\cdot} = \sum_{i=1}^{m} h_{1i}$
a_2	h_{21}	$\cdots$	h_{2m}	$h_{2\cdot} = \sum_{i=1}^{m} h_{2i}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	
a_1	h_{k1}	$\cdots$	h_{km}	$h_{k\cdot} = \sum_{i=1}^{m} h_{ki}$
	$h_{\cdot 1}$		$h_{\cdot m}$	n

Die Zeilensummen $h_{i\cdot}$ werden auch als Randhäufigkeiten des Merkmals X bezeichnet. Diese Werte sind die einfachen Häufigkeiten mit denen das Merkmal X die Werte $a_1, \dots, a_k$ annimmt, wenn Y nicht berücksichtigt wird.
Analog dazu sind die Spaltensummen die Häufigkeiten von Y unter Vernachlässigung des Merkmals X.

Kontingenztafel der relativen Häufigkeiten Da Anteile beziehungsweise Prozente häufig anschaulicher sind als absolute Häufigkeitswerte betrachtet man häufig auch die Häufigkeitstafel der relativen Häufigkeiten. Diese entsteht durch teilen durch die Gesamtzahl n.

	b_1	$\cdots$	b_m	
a_1	h_{11}	$\cdots$	h_{1m}	$f_{1\cdot} = \sum_{i=1}^{m} f_{1i}$
a_2	f_{21}	$\cdots$	f_{2m}	$f_{2\cdot} = \sum_{i=1}^{m} f_{2i}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	
a_1	f_{k1}	$\cdots$	f_{km}	$f_{k\cdot} = \sum_{i=1}^{m} f_{ki}$
	$f_{\cdot 1}$		$f_{\cdot m}$	1

3.2 Bedingte Häufigkeiten

Aus den gemeinsamen Häufigkeiten lässt sich nicht direkt auf den Zusammenhang zweier Merkmale schließen. So kann man ein Merkmal fest wählen und dann die Häufigkeitsverteilung des anderen

Merkmals unabhängig davon betrachten, mit dieser Herangehensweise kommt man zu den bedingten Häufigkeiten.

Definition 3.1: Bedingte Häufigkeiten

Für zwei Merkmale X und Y mit den Ausprägungen $a_1, \ldots, a_k$ und $b_1, \ldots, b_m$ ist

$$f_Y(b_j|a_i) = \frac{h_{ij}}{h_i}$$

die Häufigkeit des Merkmals b_j aus Y unter der Bedingung $X = a_i$.
Daraus geht durch die Werte

$$f_Y(b_1|a_i),\ \ldots,\ f_Y(b_m|a_i)$$

die bedingte Häufigkeitsverteilung von Y unter der Bedingung $X = a_i$ (Kurzschreibweise: $(Y|X = a_i)$) hervor.
Analog für eine fest gewählte Ausprägung $Y = b_i$ die bedingte Häufigkeitsverteilung $f_X(a_j|b_i)$.

3.3 Zusammenhangsmaße

Wir betracten zunächst wie sich zwei Merkmale zueinander verhalten würden, wenn keinerlei Zusammenhang zwischen ihnen bestünde.
In einer Kontingenztafel müssten sich dann die einzelnen Spalten proportional zu den Spaltensummen verhalten und analog die Zeilen zu den Zeilensummen. Daraus ergibt sich die *erwartete Häufigkeit einer Ausprägung bei unabhängigen Merkmalen* X und Y

$$\tilde{h}_{ij} = \frac{h_{i\cdot} \cdot h_{\cdot j}}{n}$$

Um den Zusammenhang zweier Merkmale zu untersuchen betrachten wir also den Unterschied zwischen den tatsächlichen Häufigkeiten h_{ij} und den jeweils erwarteten Häufigkeiten $\tilde{h}_{ij}$.

3.3.1 χ^2-Koeffizient

Basierend auf der oben angesprochenen Differenz wird das erste Zusammenhangsmaß konstruiert.

Definition 3.2: χ^2-Koeffizient

Für zwei Merkmale X und Y mit den Ausprägungen $a_1, \ldots, a_k$ und $b_1, \ldots, b_m$ ist

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} \quad \chi^2 \in [0, \infty)$$

Ist χ^2 groß, weichen die Häufigkeiten also stark von der Erwartung ab, hängen die Merkmale vermutlich voneinander ab. Sind die Abweichungen allerdings relativ klein und damit χ^2 ebenfalls, so sind die Merkmale wahrscheinlich unabhängig. Selbst bei unabhängigen Merkmalen ist meist wegen zufälligem Rauschen $\chi^2 \neq 0$, eine Entscheidung ist so also nicht möglich.

3.3.2 Kontingenzkoeffizient

Problematisch am χ^2-Koeffizienten ist die Abhängigkeit von der Dimension der Tafel. Es kann nicht ohne weiteres aus dem Wert des Koeffizienten auf eine Unabhängigkeit der Merkmale geschlossen werden. Als erster Normierungsschritt folgt daraus der Kontingenzkoeffizient.

Definition 3.3: Kontingenzkoeffizient

Aus dem χ^2-Koeffizienten für zwei Merkmale mit der Urliste der Größe n ist der Kontingenzkoeffizient

$$K = \sqrt{\frac{\chi^2}{n + \chi^2}} \quad K \in [0, K_{\max}] \text{ für } K_{\max} = \sqrt{\frac{M-1}{M}}, \quad M = \max\{m, k\}$$

Dabei seien m und k die Mächtigkeit der Ausprägungsliste der beiden Merkmale.

Korrigierter Kontingenzkoeffizient Da auch dieser für einen sinnvollen Vergleich nicht ausreichend normiert ist, wird der korrigierte Kontingenzkoeffizient eingeführt, wobei K durch $K_{\max}$ normiert wird.

$$K^* = \frac{K}{K_{\max}} \quad K^* \in [0, 1]$$

Sowohl der Kontingenzkoeffizient als auch der χ^2-Koeffizient stellen nur die Stärke des Zusammenhangs dar, nicht aber eine Richtung der Wirkungsweise.

Literaturverzeichnis

[1] L. Fahrmeir, *Statistik: Der Weg zur Datenanalyse*, Springer-Lehrbuch : SpringerLink : Bücher, Springer Spektrum, Berlin, Heidelberg, 8. aufl. 2016 ed., 2016.

[2] N. Henze, *Stochastik für Einsteiger: Eine Einführung in die faszinierende Welt des Zufalls*, SpringerLink : Bücher, Springer Spektrum, Wiesbaden, 10., überarb. aufl. 2013 ed., 2013.

[3] U. Krengel, *Einführung in die Wahrscheinlichkeitstheorie und Statistik*, Vieweg Studium, Aufbaukurs Mathematik : SpringerLink : Bücher, Vieweg+Teubner Verlag, Wiesbaden, 8., erweiterte auflage ed., 2005.